

# The high resolution vector quantization problem with Orlicz norm distortion

by

S. Dereich and C. Vormoor

*Fachbereich Mathematik und Informatik  
Philipps-Universität Marburg  
Hans-Meerwein Straße  
D-35032 Marburg*

*E-mail: dereich@mathematik.uni-marburg.de, cvormoor@skandia.de*

**Summary.** We derive a high-resolution formula for the quantization problem under Orlicz norm distortion. In this setting, the optimal point density solves a variational problem which comprises a function  $g : \mathbb{R}_+ \rightarrow [0, \infty)$  characterizing the quantization complexity of the underlying Orlicz space. Moreover, asymptotically optimal codebooks induce a tight sequence of empirical measures. The set of possible accumulation points is characterized and in most cases it consists of a single element. In that case, we find convergence as in the classical setting.

**Keywords.** Complexity; discrete approximation, high-resolution quantization; self similarity.

**2000 Mathematics Subject Classification.** 60E99, 68P30, 94A29.

## 1 Introduction

For  $d \in \mathbb{N}$ , consider an  $\mathbb{R}^d$ -valued random vector  $X$  (*the original*) defined on some probability space  $(\Omega, \mathcal{A}, \mathbb{P})$ , and denote by  $\mu = \mathcal{L}(X)$  the law of  $X$ .

We consider the quantization problem, that is for a given natural number  $N \in \mathbb{N}$  and a loss function  $\rho : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$  we ask for a codebook  $\mathcal{C} \subset \mathbb{R}^d$  consisting of at most  $N$  elements and minimizing the average loss

$$\mathbb{E}\rho(X, \mathcal{C}),$$

where  $\rho(x, A) = \inf_{y \in A} \rho(x, y)$ , for all  $x \in \mathbb{R}^d$  and  $A \subset \mathbb{R}^d$ . The quantization problem arises naturally when discretizing analog signals, and it first gained practical importance in the context of pulse-code-modulation. Research on it started in the 1940's and one finds numerous articles dedicated to the study of this problem in the engineering literature. For an overview on these developments, one may consult Gray and Neuhoﬀ (1998) (see also Cover and Thomas (1991) and Gersho and Gray (1992)). The quantization problem is also related to numerical integration Pagès, Pham and Printems (2004), and, more recently, the mathematical community became attracted by the field. In the last years a number of new publications appeared treating finite dimensional as well as infinite dimensional signals (see for instance Graf and Luschgy (2000), Gruber (2004) for vector quantization and Luschgy and Pagès (2004), Dereich et al. (2003) for functional quantization).

In this article, we consider asymptotic properties of the quantization problem when the size  $N$  of the codebook tends to infinity, the *high-resolution quantization problem*. First asymptotic formulae for vector quantization were found by Zador (1966, 1982) and Bucklew and Wise (1982).

In the classical setting (*norm based distortion*), one considers

$$\rho(x, y) = \|x - y\|^p$$

for some norm  $\|\cdot\|$  on  $\mathbb{R}^d$  and a moment  $p \geq 1$ . As long as the distribution  $\mu := \mathcal{L}(X)$  has thin tails (in an appropriate sense) one can describe asymptotically optimal codebooks via an optimal point density function: the empirical measures associated to optimal codebooks  $\mathcal{C}(N)$  of size  $N$

$$\frac{1}{N} \sum_{\hat{x} \in \mathcal{C}} \delta_{\hat{x}}$$

converge to a continuous probability measure that has density proportional to  $\left(\frac{d\mu_c}{d\lambda^d}\right)^{d/(d+p)}$ . The density will be called *optimal point density*. Here and elsewhere,  $\lambda^d$  denotes  $d$ -dimensional Lebesgue measure and  $\mu_c$  denotes the absolutely continuous part of  $\mu$  w.r.t.  $\lambda^d$ . When considering point densities we will always assume that  $\mu_c$  does not vanish. Optimal codebooks for the uniform distribution can then be used to define asymptotically optimal codebooks for general  $X$ : roughly speaking, one partitions the space  $\mathbb{R}^d$  into appropriate cubes and chooses in each cube an optimal codebook for the uniform distribution of appropriate size (according to the optimal point density). In particular, the non-continuous part of  $\mu$  has no effect on the asymptotic problem. The concept of a point density will play a crucial role in the following discussion. Its importance in the classical setting was first conjectured by Lloyd and Gersho (see Gersho (1979)). First rigorous proofs are due to Bucklew (1984). For a recent account on the theory of high resolution quantization and point density functions one may consult the monograph by Graf and Luschgy (2000).

Nowadays, the asymptotic quantization problem is well understood for loss functions  $\rho$  that are shift invariant and look locally like a power of a norm based distance, that is

$$\rho(x, y) = \rho(x - y) \quad \text{and} \quad \rho(x) = \|x\|^p + o(\|x\|^p) \quad \text{as } x \rightarrow 0$$

for some norm  $\|\cdot\|$  on  $\mathbb{R}^d$  and a power  $p \geq 1$  (see Delattre et al. (2004)). Here and thereafter  $o$  and  $\mathcal{O}$  denote the Landau symbols. For all these distortion measures one

regains the same optimal point density as for the corresponding norm based distortion measures. Since the optimal point density only depends on the behavior of  $\rho$  close to 0, quantization schemes based on this density may show bad performance for moderate  $N$ . This will occur, when for the original  $X$  and the approximation  $\hat{X}$  the distances  $\rho(X, \hat{X})$  and  $\|X - \hat{X}\|^p$  differ significantly.

As a generalization of the above setting, we suggest to use Orlicz norms as a measure for the loss inferred when approximating the original  $X$  by the closest point in a codebook  $\mathcal{C}$ . One reason for this is that optimal codebooks for general distortions are also optimal codebooks for certain Orlicz norm distortions. For moderate  $N$  the optimal point density of the corresponding Orlicz norm distortion seems to be the favorable choice as basis for the construction of good codebooks since it incorporates  $\rho$  as a whole and not only its asymptotics in zero.

Let us introduce the main notation. Let  $\varphi : [0, \infty) \rightarrow [0, \infty)$  be an increasing, left continuous function with  $\lim_{t \downarrow 0} \varphi(t) = 0$ . Note that this implies that  $\varphi$  is lower semicontinuous. We assume that  $\varphi \neq 0$ , let  $E = (\mathbb{R}^d, \|\cdot\|)$  denote an arbitrary Banach space, and denote by  $d(\cdot, \cdot)$  the associated distance. For any  $\mathbb{R}^d$ -valued r.v.  $Z$ , the Orlicz norm  $\|\cdot\|_\varphi$  is defined as

$$\|Z\|_\varphi = \inf \left\{ t \geq 0 : \mathbb{E} \varphi \left( \frac{\|Z\|}{t} \right) \leq 1 \right\},$$

with the convention that the infimum of the empty set is equal to infinity. Actually, the left continuity of  $\varphi$  together with monotone convergence imply that the infimum is attained, whenever the set is nonempty. We set

$$L^\varphi(\mathbb{P}) = \{Z : Z \text{ } \mathbb{R}^d\text{-valued r.v. with } \|Z\|_\varphi < \infty\}.$$

Note that  $\|\cdot\|_\varphi$  defines a norm on  $L^\varphi(\mathbb{P})$  when  $\varphi$  is convex, whereas otherwise, the triangle inequality does not hold. For our analysis we do not require that  $\varphi$  be convex. Nevertheless, with slight misuse of notation, we will allow ourselves to call  $\|\cdot\|_\varphi$  an Orlicz norm. Choosing  $\varphi(t) = t^p$ ,  $p \geq 1$ , yields the usual  $L^p(\mathbb{P})$ -norm, which will be denoted by  $\|\cdot\|_p$ .

For  $N \geq 1$ , we consider the *quantization error* given by

$$\delta(N|X, \varphi) = \inf_{\hat{X}} \|X - \hat{X}\|_\varphi,$$

where the infimum is taken over all r.v.'s  $\hat{X}$  (*reconstructions*) satisfying the range constraint  $|\text{range}(\hat{X})| \leq N$ . In the case where  $\varphi(t) = t^p$  for a  $p > 0$ , we write briefly  $\delta(N|X, p)$  for the corresponding quantization error.

Let us compare the Orlicz norm distortion with the classical setting. Suppose that  $\hat{X}$  is an optimal  $N$ -point quantizer under the distortion  $\rho(x, y) = f(\|x - y\|)$ , where  $f$  is a left continuous and strictly increasing function with  $f(0) = \lim_{t \downarrow 0} f(t) = 0$ . Then one can easily verify that  $\hat{X}$  is also an optimal  $N$ -point quantizer in the Orlicz norm setting when choosing  $\varphi(t) = f(t)/\Delta$  and  $\Delta = \mathbb{E}[\rho(X, \hat{X})]$ . Thus optimal quantizers in the classical setting correspond to optimal quantizers in the Orlicz-norm setting. As we will see later, in most cases each choice of  $\Delta$  (or  $\varphi$ ) leads to a unique optimal point density function. Thus the optimal point density function in the classical setting is replaced by a whole

family of densities: good descriptions are now obtained by choosing the parameter  $\Delta$  accordingly. We believe that the optimal Orlicz point density is a favourable description of good codebooks for moderate  $N$ . Let us illustrate this in the case where  $f(t) = \varphi(t) = \exp(t) - 1$  (meaning that  $\Delta = 1$ ). Whereas the Orlicz norm point density attains rather large values even at points where the density  $\frac{d\mu_c}{d\lambda^d}$  is very small (see Example 1.4), the classical setting neglects the growth of  $f$  and one retrieves the optimal density of the norm based distortion to the power 1.

In our notation, the asymptotics under the classical  $L^p$ -norm distortion reads as follows (see Graf and Luschgy (2000)): Let  $p \geq 1$ ,  $U$  denote a uniformly distributed r.v. on  $[0, 1]^d$ , and set

$$q(E, p) = \inf_{N \geq 1} N^{1/d} \delta(N|U, p).$$

If for some  $\tilde{p} > p$ ,  $\mathbb{E}\|X\|^{\tilde{p}} < \infty$  (*concentration assumption*), then

$$\lim_{N \rightarrow \infty} N^{1/d} \delta(N|X, p) = q(E, p) \left\| \frac{d\mu_c}{d\lambda^d} \right\|_{L^{d/(d+p)}(\mathbb{R}^d)}^{1/p}, \quad (1)$$

where  $\mu_c$  denotes the absolutely continuous part of  $\mu$ .

The analysis of the Orlicz norm setting is based on the concept of a point allocation density. We shall see that an optimal point allocation density is given as a minimizer of a variational problem. Unfortunately, in this context the minimization problem cannot be solved in closed form. We will prove the existence and a dual characterization of the solution. The quantity  $q(E, p)$  corresponds in the general setting to a convex decreasing function  $g : (0, \infty) \rightarrow [0, \infty)$  which may be defined via

$$g(\eta) = \lim_{N \rightarrow \infty} \inf_{\mathcal{C}(N)} \mathbb{E} \varphi((N/\eta)^{1/d} d(U, \mathcal{C}(N))), \quad (2)$$

where the infima are taken over all finite sets  $\mathcal{C}(N) \subset \mathbb{R}^d$  with  $|\mathcal{C}(N)| \leq N$  and  $U$  denotes a uniformly distributed r.v. on the unit cube  $[0, 1]^d$ . Moreover, we set  $g(0) = \liminf_{\eta \downarrow 0} g(\eta) \in [0, \infty]$ . The function  $g$  depends on the Banach space and the Orlicz norm. It will be analyzed in Section 2.  $g$  can be represented as a particular integral in the situations where  $E = l_\infty^d$  or  $E = l_2^2$ . If  $\varphi$  induces the  $L^p(\mathbb{P})$ -norm (i.e.  $\varphi(t) = t^p$ ), then due to (1)

$$g(\eta) = q(E, p)^p \eta^{-p/d}.$$

If the measure  $\mu$  is not compactly supported, our analysis relies on a concentration property. As in the  $L^p(\mathbb{P})$ -setting, this can be done by assuming the finiteness of an integral  $\mathbb{E}\Psi(\|X\|)$  for some function  $\Psi$  satisfying a growth condition (Condition (G), see Definition 7.2). Let us state the main theorem.

**Theorem 1.1.** *Assume that  $\Psi$  satisfies the growth condition (G) and that  $\mathbb{E}\Psi(\|X\|) < \infty$ . Then*

$$\lim_{N \rightarrow \infty} N^{1/d} \delta(N|X, \varphi) = I^{1/d},$$

where  $I$  is the finite minimal value in the point allocation problem. It is given by

$$I = \inf_{\xi} \int \xi(x) dx, \quad (3)$$

where the infimum is taken over all non-negative Lebesgue integrable functions  $\xi$  with

$$\int_{\mathbb{R}^d} g(\xi(x)) d\mu_c(x) \leq 1.$$

Alternatively, one can represent  $I$  by the dual formula

$$I = \sup_{\kappa > 0} \frac{1}{\kappa} \left( \int_{\mathbb{R}^d} \bar{g}\left(\frac{\kappa}{h(x)}\right) d\mu_c(x) - 1 \right),$$

where  $\bar{g}(t) = \inf_{\eta > 0} [g(\eta) + \eta t]$  and  $h(x) = \frac{d\mu_c}{d\lambda^d}(x)$ . Moreover, one has  $I > 0$  if and only if

$$\mu_c(\mathbb{R}^d) \sup_{t \geq 0} \varphi(t) > 1,$$

where we use the convention that  $0 \cdot \infty = 0$ .

**Remark 1.2.** • If we choose  $\varphi(t) = t^p$  in the former theorem, we obtain the classical result, since  $\Psi(t) = t^q$  with  $q > p$  satisfies the growth condition (G), see Example 7.3 for an even weaker assumption.

- The point allocation problem is studied in Section 5. In particular, representations for optimizers can be found in Theorem 5.1 and Remark 5.2.

We keep  $I$  as the minimal value in the point allocation problem given by (3). If  $I$  is strictly bigger than 0, we denote by  $\mathcal{M}$  the set of probability measures on the Borel sets of  $\mathbb{R}^d$  associated to the minimizers of the point allocation problem, i.e.,

$$\mathcal{M} = \left\{ \nu : \frac{d\nu}{d\lambda^d} = \bar{\xi}, \int_{\mathbb{R}^d} g(I \bar{\xi}(x)) d\mu_c(x) = 1, \int_{\mathbb{R}^d} \bar{\xi}(x) dx = 1 \right\}. \quad (4)$$

**Theorem 1.3.** Assume that  $I \in (0, \infty)$  and denote by  $\mathcal{C}(N)$ ,  $N \in \mathbb{N}$ , asymptotically optimal codebooks of size  $N$ , that is

$$\limsup_{N \rightarrow \infty} N^{1/d} \|d(X, \mathcal{C}(N))\|_\varphi \leq I^{1/d}.$$

Then the empirical measures  $\nu_N$  given by

$$\nu_N = \frac{1}{N} \sum_{\hat{x} \in \mathcal{C}(N)} \delta_{\hat{x}}$$

form a tight sequence of probability measures, and any accumulation point of  $(\nu_N)_{N \in \mathbb{N}}$  lies in  $\mathcal{M}$ . If  $g$  is strictly convex, then the set  $\mathcal{M}$  contains exactly one measure  $\nu$ , and

$$\lim_{N \rightarrow \infty} \nu_N = \nu \quad \text{weakly.}$$

As an example we present implications of our results for the standard normal distribution under a particular  $\varphi$  growing exponentially fast:

**Example 1.4.** Let  $\mu$  denote the standard normal distribution,  $(E, \|\cdot\|) = (\mathbb{R}, |\cdot|)$  and  $\varphi : [0, \infty) \rightarrow [0, \infty)$ ,  $x \mapsto \exp(x) - 1$ . Then  $\Psi(x) = \exp(x^{3/2})$  satisfies the growth condition (G) (see Example 7.3), and Theorem 1.1 is thus applicable. Moreover, (see Remark 2.2 and Lemma 2.3)

$$g(\eta) = 2 \int_0^{1/2} \varphi(t/\eta) dt = \eta e^{1/(2\eta)} - 2\eta - 1.$$

Note that  $g$  is strictly convex, so that there has to exist a unique normalized optimal point density  $\bar{\xi}$ . In order to approximate the optimal value for  $\kappa$ , we used numerical methods to obtain  $\kappa \approx 0.699$ . Due to (23) and (24), the optimal point density is given by

$$\xi(x) = (-g')^{-1}\left(\frac{\kappa}{h(x)}\right), \quad x \in \mathbb{R},$$

and  $I = \int_{\mathbb{R}} \xi(x) dx \approx 2.88$ . Next, elementary calculus gives

$$(-g')^{-1}(t) = \frac{1}{2 \log(t) - \log(2 \log t) + o(1)} \quad \text{as } t \rightarrow \infty$$

so that the normalized point density  $\bar{\xi} = \xi/I$  satisfies

$$\bar{\xi}(x) \sim \frac{1}{I} \frac{1}{x^2} \quad \text{as } |x| \rightarrow \infty.$$

Hence,  $\bar{\xi}$  decays to zero much more slowly than in the classical setting.

The article is outlined as follows. In Section 2, we begin with an analysis of the function  $g$ . In Section 3 we construct asymptotically good codebooks based on a given point allocation measure. Up to this stage, we are restricting ourselves to absolutely continuous measures with compact support. In Section 4, we turn things around and prove a lower bound based on a given point density measure. This bound implies the lower bound in Theorem 1.1 and proves a part of Theorem 1.3. The estimates of Sections 3 and 4 lead to the variational problem characterising the point density, and this is treated in Section 5. In the last two sections, we treat the upper bounds in the quantization problem for singular and non-compactly supported measures. In particular, we derive a concentration analog which guarantees that the quantization error is of order  $\mathcal{O}(N^{-1/d})$ . Finally, we combine the estimates and prove the general upper bound in Theorem 1.1.

It is convenient to use the symbols  $\sim$ ,  $\lesssim$  and  $\approx$ . We write  $f \sim g$  iff  $\lim \frac{f}{g} = 1$ , while  $f \lesssim g$  stands for  $\limsup \frac{f}{g} \leq 1$ . Finally,  $f \approx g$  means  $0 < \liminf \frac{f}{g} \leq \limsup \frac{f}{g} < \infty$ .

## 2 First estimates for the uniform distribution

In this section,  $X$  denotes a uniformly distributed r.v. on  $[0, 1)^d$ . For  $\eta > 0$  and  $N \geq 1$ , we consider

$$f_N(\eta) := \inf_{\mathcal{C}(N)} \mathbb{E} \varphi\left((N/\eta)^{1/d} \min_{\hat{x} \in \mathcal{C}(N)} \|X - \hat{x}\|\right), \quad (5)$$

where the infimum is taken over all codebooks  $\mathcal{C}(N) \subset \mathbb{R}^d$  of size  $\lfloor N \rfloor$ . Here and elsewhere,  $\lfloor N \rfloor$  denotes the largest integer smaller or equal to  $N$ . By a straightforward argument, the lower semicontinuity of  $\varphi$  implies that the function

$$(\mathbb{R}^d)^{\lfloor N \rfloor} \ni (\hat{x}_1, \dots, \hat{x}_{\lfloor N \rfloor}) \mapsto \mathbb{E} \varphi \left( (N/\eta)^{1/d} \min_{i=1, \dots, \lfloor N \rfloor} \|X - \hat{x}_i\| \right) \in [0, \infty)$$

is also lower semicontinuous. In the minimization problem (5), it suffices to allow for codebook entries that are elements of a sufficiently large compact set. So the lower semicontinuity implies the existence of an optimal codebook. We usually denote by  $\hat{X}^{(N)}$  or  $\hat{X}^{(N, \eta)}$  an optimal reconstruction attaining at most  $N$  different values, that is  $\hat{X} = \hat{X}^{(N)}$  is a minimizer of

$$\mathbb{E} \varphi \left( (N/\eta)^{1/d} \|X - \hat{X}\| \right),$$

among all r.v.'s satisfying the range constraint  $|\text{range}(\hat{X})| \leq N$ . Now define the function  $g$  by

$$g(\eta) = \inf_{N \geq 1} f_N(\eta), \quad \eta > 0.$$

We start with a derivation of the structural properties of  $g$ . In particular, we show the validity of (2).

**Theorem 2.1.** *The function  $g : \mathbb{R}_+ \rightarrow [0, \infty)$  is decreasing and convex, and satisfies  $\lim_{\eta \rightarrow \infty} g(\eta) = 0$  and  $\lim_{\eta \downarrow 0} g(\eta) = \sup_{t \geq 0} \varphi(t)$ . Moreover, for  $\eta > 0$ ,*

$$\lim_{N \rightarrow \infty} \mathbb{E} \varphi \left( (N/\eta)^{1/d} \|X - \hat{X}^{(N)}\| \right) = g(\eta). \quad (6)$$

We will sometimes use the convention  $g(0) = \lim_{\eta \downarrow 0} g(\eta)$ . Note that  $g(0)$  is finite iff  $\varphi$  is bounded. Moreover, we set  $f_N(\eta) = \infty$  for  $N \in [0, 1)$ .

**Remark 2.2.** In general, computing the function  $g$  explicitly constitutes a hard problem. However, as for the classical  $L^q$ -norm distortion, one can calculate  $g$  when  $E = \mathbb{R}^d$  is endowed with supremum-norm, and in the case where  $E$  is the two dimensional Euclidean space. In such cases, the same lattice quantizers can be used to construct asymptotically optimal codebooks and to compute the function  $g$ . The case of the supremum-norm is trivial since the unit ball is space filling.

**Lemma 2.3.** *Let  $U$  be uniformly distributed on a centered regular hexagon  $V$  in  $\mathbb{R}^2$  having unit area, and assume that  $E$  is the 2-dimensional Euclidean space. One has*

$$g(\eta) = \mathbb{E} \varphi \left( \eta^{-1/2} \|U\| \right).$$

The proof is similar as in the classical setting and therefore omitted, see (Graf and Luschgy, 2000, Theorem 8.15) and Fejes Tóth (1972).

In order to prove Theorem 2.1, we use the inequality below. It is essentially a consequence of the self similarity of  $X$ .

**Proposition 2.4.** *Let  $M \in \{k^d : k \in \mathbb{N}\}$ ,  $\eta_2 > \eta_1 > 0$  and let  $\eta = \alpha\eta_1 + \beta\eta_2$  be a convex combination of  $\eta_1$  and  $\eta_2$ . Then for any  $N \geq 1$  one has*

$$f_N(\eta) \leq a(M) f_{N_1/M}(\eta_1) + b(M) f_{N_2/M}(\eta_2), \quad (7)$$

where

- $N_1 = \frac{\eta_1}{\eta}N$  and  $N_2 = \frac{\eta_2}{\eta}N$ ,
- $b(M) = \lfloor \beta M \rfloor / M$  and  $a(M) = 1 - b(M)$ .

Additionally, one has for  $N \geq 1$  and  $\eta > 0$ ,

$$f_N(\eta) \leq f_{N/M}(\eta). \quad (8)$$

**Proof.** Fix  $N \in \mathbb{N}$  and let  $\eta_1, \eta_2, \alpha, \beta, N_1, N_2, M = k^d$  be as in the proposition. Let  $C_1 = [0, 1/k)^d$ . We decompose the cube  $[0, 1)^d$  into an appropriate union  $\bigcup_{i=1}^M C_i$  of disjoint sets  $C_1, \dots, C_M$ , where each set  $C_2, \dots, C_M$  is a translate of  $C_1$ . Moreover, let  $X_i$  denote a uniformly distributed r.v. on  $C_i$ . Since  $\mathcal{U}([0, 1)^d) = \frac{1}{M} \sum_{i=1}^M \mathcal{U}(C_i)$ , one has, in analogy to Lemma 4.14 in Graf and Luschgy (2000),

$$f_N(\eta) = \mathbb{E} \varphi\left((N/\eta)^{1/d} \|X - \hat{X}^{(N)}\|\right) \leq \frac{1}{M} \sum_{i=1}^M \mathbb{E} \varphi\left((N/\eta)^{1/d} \|X_i - \hat{X}_i^{(\tilde{N}_i)}\|\right)$$

for any  $[1, \infty)$ -valued sequence  $(\tilde{N}_i)_{i=1, \dots, M}$  with  $\sum_i \tilde{N}_i \leq N$ . Here,  $\hat{X}_i^{(\tilde{N}_i)}$  denotes an optimal quantizer for  $X_i$  among all quantizers attaining at most  $\tilde{N}_i$  different values. Since the distributions  $\mathcal{U}(C_i)$  can be transformed into  $\mathcal{U}(C_1)$  through a translation, one obtains

$$f_N(\eta) \leq \frac{1}{M} \sum_{i=1}^M \mathbb{E} \varphi\left((N/\eta)^{1/d} \|X_1 - \hat{X}_1^{(\tilde{N}_i)}\|\right).$$

Self similarity ( $\mathcal{L}(X_1) = \mathcal{L}(\frac{1}{k}X)$ ) then implies that

$$f_N(\eta) \leq \frac{1}{M} \sum_{i=1}^M \mathbb{E} \varphi\left((N/\eta)^{1/d} \frac{1}{k} \|X - \hat{X}^{(\tilde{N}_i)}\|\right).$$

Note that the assertion of the proposition is trivial if  $N_1/M < 1$ , so that we may assume  $N_2/M \geq N_1/M \geq 1$ . We now choose for  $\lfloor \beta M \rfloor$  indices  $\tilde{N}_i = N_2/M$  and for  $M - \lfloor \beta M \rfloor$  indices  $\tilde{N}_i = N_1/M$ . Then  $\sum_i \tilde{N}_i \leq N$ , so that

$$\begin{aligned} f_N(\eta) &\leq a(M) \mathbb{E} \varphi\left((N/\eta)^{1/d} \frac{1}{k} \|X - \hat{X}^{(N_1/M)}\|\right) \\ &\quad + b(M) \mathbb{E} \varphi\left((N/\eta)^{1/d} \frac{1}{k} \|X - \hat{X}^{(N_2/M)}\|\right) \\ &= a(M) f_{N_1/M}(\eta_1) + b(M) f_{N_2/M}(\eta_2), \end{aligned}$$

where  $b(M) = \lfloor \beta M \rfloor / M$  and  $a(M) = 1 - b(M)$ . Analogously, setting  $\tilde{N}_i = N/M$  for  $i = 1, \dots, M$ , we obtain that  $f_N(\eta) \leq f_{N/M}(\eta)$ .  $\square$

**Proof of Theorem 2.1.** Obviously  $g$  is decreasing. First we prove that for arbitrary  $\eta > 0$ ,

$$g(\eta) \leq \limsup_{N \rightarrow \infty} f_N(\eta) \leq g_-(\eta). \quad (9)$$

Fix  $\varepsilon > 0$  and choose  $\eta_0 \in (0, \eta)$  so that  $g(\eta_0) \leq g_-(\eta) + \varepsilon/2$ . Moreover, fix  $N_0 \geq 1$  with  $f_{N_0}(\eta) \leq g(\eta_0) + \varepsilon/2$ . For  $N \geq N_0$ , we decompose  $N$  into  $N = N_0 k^d + \tilde{N}$ , where  $k = k(N) \in \mathbb{N}$  and  $\tilde{N} = \tilde{N}(N) \in \mathbb{N}_0$  are chosen so that  $N < (k+1)^d N_0$ . Then

$$\begin{aligned} f_N(\eta) &= \mathbb{E} \varphi\left((N/\eta)^{1/d} \|X - \hat{X}^{(N)}\|\right) \\ &\leq \mathbb{E} \varphi\left((N_0 k^d / (N_0 k^d \eta / N))^{1/d} \|X - \hat{X}^{(N_0 k^d)}\|\right) = f_{N_0 k^d}(\eta N_0 k^d / N), \end{aligned}$$

and inequality (8) implies that for  $M = M(N) = k^d$ :

$$f_N(\eta) \leq f_{N_0}(\eta N_0 M / N).$$

Note that  $N_0 k^d \leq N < N_0(k+1)^d$ , hence:  $\lim_{N \rightarrow \infty} \eta N_0 k^d / N = \eta$ . Consequently, there exists  $N_1 \geq N_0$  such that for all  $N \geq N_1$  one has:  $\eta N_0 M / N \geq \eta_0$ , and

$$f_N(\eta) \leq f_{N_0}(\eta N_0 M / N) \leq f_{N_0}(\eta_0) \leq g_-(\eta) + \varepsilon$$

for all  $N \geq N_1$ . Since  $\varepsilon > 0$  was arbitrary statement (9) follows.

We now prove that  $g_-$  is convex. Let  $\eta_2 > \eta_1 > 0$  and let  $\eta = \alpha \eta_1 + \beta \eta_2$  be a convex combination of  $\eta_1$  and  $\eta_2$  and suppose that  $g_-(\eta_1)$  is finite. Fix  $k \in \mathbb{N}$  and let  $M = k^d$ ,  $a(M)$  and  $b(M)$  be as in Proposition 2.4. Moreover, for given  $N \in \mathbb{N}$  we let  $N_1 = N_1(N)$  and  $N_2 = N_2(N)$  be as in the previous proposition. Then inequality (7) implies that

$$f_N(\eta) \leq a(M) f_{N_1/M}(\eta_1) + b(M) f_{N_2/M}(\eta_2).$$

Therefore, formula (9) and the left continuity of  $g_-$  give

$$g(\eta) \leq \limsup_{N \rightarrow \infty} f_N(\eta) \leq a(M) g_-(\eta_1) + b(M) g_-(\eta_2).$$

Recall that  $M \in \{k^d : k \in \mathbb{N}\}$  was arbitrary. Since  $\lim_{M \rightarrow \infty} a(M) = \alpha$  and  $\lim_{M \rightarrow \infty} b(M) = \beta$  we conclude that

$$g(\eta) \leq \alpha g_-(\eta_1) + \beta g_-(\eta_2).$$

For the general statement, observe that

$$g_-(\eta) = \lim_{\delta \downarrow 0} g(\eta - \delta) \leq \limsup_{\delta \downarrow 0} [\alpha g_-(\eta_1 - \delta) + \beta g_-(\eta_2 - \delta)] = \alpha g_-(\eta_1) + \beta g_-(\eta_2).$$

Consequently,  $g_-$  is convex, and a fortiori it is continuous. Therefore, the functions  $g$  and  $g_-$  coincide, which proves (6).

It remains to prove the asymptotic statements for  $g$ . First note that

$$g(\eta) \leq f_1(\eta) \leq \mathbb{E} \varphi(\eta^{-1/d} \|X\|) \leq \varphi(\eta^{-1/d} \sup_{x \in [0,1]^d} \|x\|) \longrightarrow 0$$

as  $\eta \rightarrow \infty$ . On the other hand, one has for  $\eta > 0$ ,  $\varepsilon > 0$  and  $N \in \mathbb{N}$ ,

$$\begin{aligned} f_N(\eta) &= \mathbb{E} \varphi\left((N/\eta)^{1/d} \|X - \hat{X}^{(N)}\|\right) \\ &\geq \mathbb{E} \left[ 1_{\{\|X - \hat{X}^{(N)}\| \geq \varepsilon/N^{1/d}\}} \varphi\left((N/\eta)^{1/d} \|X - \hat{X}^{(N)}\|\right) \right] \\ &\geq (1 - N \lambda^d(B(0, \varepsilon/N^{1/d}))) \varphi(\varepsilon/\eta^{1/d}) \\ &= (1 - \lambda^d(B(0, \varepsilon))) \varphi(\varepsilon/\eta^{1/d}), \end{aligned}$$

so that

$$g(\eta) \geq (1 - \lambda^d(B(0, \varepsilon))) \varphi(\varepsilon/\eta^{1/d}) \xrightarrow[\eta \downarrow 0]{} (1 - \lambda^d(B(0, \varepsilon))) \sup_{t \geq 0} \varphi(t).$$

Since  $g(\eta) \leq \sup_{t \geq 0} \varphi(t)$  and  $\varepsilon > 0$  was arbitrary, the assertion follows.  $\square$

### 3 The upper bound (1st step)

In this section, we consider an original  $X$  with law  $\mu \ll \lambda^d$ . Moreover, we assume that  $\mu$  is compactly supported and fix  $l > 0$  large enough so that  $\mu(C) = 1$  for  $C = [-l, l]^d$ .

Based on a given integrable function  $\xi : \mathbb{R}^d \rightarrow [0, \infty)$  (*point density*), we define codebooks and control their efficiency.

**Proposition 3.1.** *There exist codebooks  $\mathcal{C}(N)$ ,  $N \geq 1$ , such that  $\lim_{N \rightarrow \infty} \frac{1}{N} |\mathcal{C}(N)| = \|\xi\|_{L^1(\mathbb{R}^d)}$  and*

$$\limsup_{N \rightarrow \infty} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) \leq \int g(\xi(x)) d\mu(x).$$

**Proof.** It suffices to prove the assertion for functions  $\xi$  that are uniformly bounded away from 0 on  $C$ . If this is not the case, one can consider  $\bar{\xi} = \xi + \varepsilon 1_C$  for some  $\varepsilon > 0$ . Then the statement says that there exist codebooks  $\mathcal{C}^\varepsilon(N)$ ,  $N \geq 1$ , with  $|\mathcal{C}^\varepsilon(N)| \sim N(\|\xi\|_{L^1(\mathbb{R}^d)} + \varepsilon \lambda^d(C))$  satisfying

$$\limsup_{N \rightarrow \infty} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}^\varepsilon(N))) \leq \int g(\bar{\xi}(x)) d\mu(x) \leq \int g(\xi(x)) d\mu(x),$$

and a diagonalization argument for  $\varepsilon \downarrow 0$  proves the general assertion.

Fix  $m \in \mathbb{N}$ , let  $C_1 = [0, l/2^m]^d$  and decompose  $C$  into a finite disjoint union

$$C = \bigcup_{i=1}^M C_i,$$

where  $M = 2^{(m+1)d}$  and  $C_2, \dots, C_M$  are translates of  $C_1$ . For  $i = 1, \dots, M$ , we denote by  $X_i$  a uniformly distributed r.v. on  $C_i$ , and let  $\mu^m = \sum_{i=1}^M \mu(C_i) \mathcal{U}(C_i)$ . Moreover, we let

$$h_m = \frac{d\mu^m}{d\lambda^d} = \sum_{i=1}^M \frac{\mu(C_i)}{\lambda^d(C_i)} \cdot 1_{C_i},$$

and denote by  $\nu$  the measure given by  $\nu(A) = \int_A \xi d\lambda^d$ ,  $A \in \mathcal{B}(\mathbb{R}^d)$ .

We introduce the codebooks of interest. For some fixed  $\kappa > 0$ , let

$$\tilde{\mathcal{C}}(N) = (\kappa N^{-1/d} \mathbb{Z}^d) \cap C, \quad N \geq 1,$$

and let  $\mathcal{C}_i(N)$  denote codebooks of size  $N_i = N_i(N) = N \nu(C_i)$  minimizing  $\mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}_i(N)))$ . We consider the efficiency of the codebooks

$$\mathcal{C}(N) = \tilde{\mathcal{C}}(N) \cup \bigcup_{i=1}^M \mathcal{C}_i(N), \quad N \geq 1.$$

First, note that

$$|\tilde{\mathcal{C}}(N)| \leq (1 + 2 \frac{l}{\kappa} N^{1/d})^d \sim \left(\frac{2l}{\kappa}\right)^d N, \quad N \rightarrow \infty,$$

hence:

$$|\mathcal{C}(N)| \lesssim \left( \|\xi\|_{L^1(\mathbb{R}^d)} + \left(\frac{2l}{\kappa}\right)^d \right) N, \quad N \rightarrow \infty. \quad (10)$$

It remains to estimate the expectation  $\mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N)))$  for large  $N \geq 1$ . Observe that  $d(x, \tilde{\mathcal{C}}(N)) \leq \kappa N^{-1/d} \sup_{x \in [0,1)^d} \|x\|$  for all  $x \in C$  so that

$$\begin{aligned} & \left| \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) - \int \varphi(N^{1/d} d(x, \mathcal{C}(N))) d\mu^m(x) \right| \\ &= \left| \int \varphi(N^{1/d} d(x, \mathcal{C}(N))) (h(x) - h_m(x)) dx \right| \leq \varphi(c\kappa) \|h - h_m\|_{L^1(\mathbb{R}^d)}, \end{aligned} \quad (11)$$

where  $c = \sup_{x \in [0,1)^d} \|x\|$  is a universal constant. Moreover,

$$\begin{aligned} \int \varphi(N^{1/d} d(x, \mathcal{C}(N))) d\mu^m(x) &= \sum_{i=1}^M \mu(C_i) \mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}(N))) \\ &\leq \sum_{i=1}^M \mu(C_i) \mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}_i(N))). \end{aligned}$$

Now let  $U$  denote a  $\mathcal{U}([0,1)^d)$ -distributed r.v. Due to the optimality assumption on the choice of  $\mathcal{C}_i(N)$ , a translation and scaling then yields

$$\mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}_i(N))) = \mathbb{E} \varphi\left(N^{1/d} \frac{l}{2^{m+1}} \|U - \hat{U}^{(N_i)}\|\right),$$

where  $\hat{U}^{(N_i)}$  denotes a reconstruction minimizing the latter expectation among all r.v. with a range of size  $N_i$ . Next, rewriting the previous expectation as

$$\mathbb{E} \varphi \left( N^{1/d} \frac{l}{2^{m+1}} \|U - \hat{U}^{(N_i)}\| \right) = f_{N_i}(\nu(C_i)/\lambda^d(C_i)),$$

it follows that

$$\int \varphi(N^{1/d} d(x, \mathcal{C}(N))) d\mu^m(x) \leq \sum_{i=1}^M \mu(C_i) f_{N_i}(\nu(C_i)/\lambda^d(C_i)).$$

As  $N \rightarrow \infty$ , every  $N_i$ ,  $i = 1, \dots, M$ , converges to  $\infty$ , and one has

$$\sum_{i=1}^M \mu(C_i) f_{N_i}(\nu(C_i)/\lambda^d(C_i)) \rightarrow \sum_{i=1}^M \mu(C_i) g\left(\frac{\nu(C_i)}{\lambda^d(C_i)}\right) = \int g(\xi_m(x)) d\mu(x),$$

where  $\xi_m = \sum_{i=1}^M \frac{\nu(C_i)}{\lambda^d(C_i)} \cdot 1_{C_i}$ . Putting everything together yields

$$\limsup_{N \rightarrow \infty} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) \leq \int g(\xi_m(x)) d\mu(x) + \kappa \|h - h_m\|_{L^1(\mathbb{R}^d)} \sup_{x \in [0,1]^d} \|x\|.$$

The function  $\xi_m$  converges to  $\xi$  as  $m \rightarrow \infty$  in  $\lambda^d$ -a.a. points  $x$  (see Cohn (1980), Theorem 6.2.3). Recall that by construction  $\xi_m$  is bounded from below on  $C$ , and hence dominated convergence gives

$$\lim_{m \rightarrow \infty} \int g(\xi_m(x)) d\mu(x) = \int g(\xi(x)) d\mu(x).$$

Analogously,  $h_m$  converges to  $h$  in  $\lambda^d$ -a.a. points  $x$  and due to Scheffé's theorem (see Billingsley (1979), Theorem 16.11)  $h_m$  converges to  $h$  in  $L^1(\mathbb{R}^d)$  as  $m \rightarrow \infty$ .

For arbitrary  $\varepsilon > 0$ , we can choose  $\kappa > 0$  sufficiently large to ensure that the size of  $\mathcal{C}(N)$  (see (10)) satisfies

$$|\mathcal{C}(N)| \lesssim (1 + \varepsilon) \|\xi\|_{L^1(\mathbb{R}^d)} N.$$

Finally, it remains to pick  $m \in \mathbb{N}$  sufficiently large so that

$$\limsup_{N \rightarrow \infty} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) \leq \int g(\xi(x)) d\mu(x) + \varepsilon,$$

and the general statement then follows from a diagonalization argument.  $\square$

## 4 The lower bound

From now on, let  $X$  be an arbitrary random vector on  $\mathbb{R}^d$  with law  $\mu$ . In this section, we change our viewpoint: for an index set  $\mathbb{I} \subset [1, \infty)$  with  $\sup \mathbb{I} = \infty$ , we consider arbitrary finite codebooks  $\mathcal{C}(N)$ ,  $N \in \mathbb{I}$ , and ask for asymptotic lower bounds of

$$\mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N)))$$

as  $N \rightarrow \infty$ . Our computations are based on the assumption that the empirical measures

$$\nu^N = \frac{1}{N} \sum_{\hat{x} \in \mathcal{C}(N)} \delta_{\hat{x}}, \quad N \in \mathbb{I},$$

associated to  $\mathcal{C}(N)$  converge vaguely to some locally finite measure  $\nu$  on  $\mathbb{R}^d$ .

**Proposition 4.1.** *Letting  $\nu_c$  denote the absolutely continuous part of  $\nu$ , one has*

$$\liminf_{N \rightarrow \infty} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) \geq \int g\left(\frac{d\nu_c}{d\lambda^d}\right) d\mu_c(x).$$

**Proof.** It suffices to prove that for an arbitrary  $l > 0$ :

$$\liminf_{N \rightarrow \infty} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) \geq \int_{[-l, l]^d} g\left(\frac{d\nu_c}{d\lambda^d}\right) d\mu_c(x).$$

Indeed, the assertion then follows immediately by monotone convergence. For a given  $m \in \mathbb{N}$ , just as in the proof of the upper bound, we decompose the set  $C = [-l, l]^d$  into a disjoint union  $C = \bigcup_{i=1}^M C_i$ , where  $M = 2^{m+1}$ ,  $C_1 = [0, l/2^m]^d$ , and  $C_2, \dots, C_M$  are translates of  $C_1$ . Again we consider the measure  $\mu^m = \sum_{i=1}^M \mu_c(C_i) \mathcal{U}(C_i)$  and the density

$$h_m = \sum_{i=1}^M \frac{\mu_c(C_i)}{\lambda^d(C_i)} \cdot 1_{C_i}.$$

Analogously, we let  $\xi_m = \sum_{i=1}^M \frac{\nu(\bar{C}_i)}{\lambda^d(\bar{C}_i)} \cdot 1_{C_i}$ . For some fixed  $\kappa > 0$ , we extend the codebooks  $\mathcal{C}(N)$  to

$$\mathcal{C}^{(1)}(N) = \mathcal{C}(N) \cup ((\kappa N^{-1/d} \mathbb{Z}^d) \cap C).$$

Then, just as in (11), one has

$$\left| \int_C \varphi(N^{1/d} d(x, \mathcal{C}^{(1)}(N))) d\mu_c(x) - \int_C \varphi(N^{1/d} d(x, \mathcal{C}^{(1)}(N))) d\mu^m(x) \right| \leq \varphi(c\kappa) \|h - h_m\|_{L^1(C)}, \quad (12)$$

where  $c = \sup_{x \in [0, 1]^d} \|x\|$ .

Next, we control the approximation efficiency of  $\mathcal{C}^{(1)}(N)$  for the measure  $\mu^m$ . We fix  $\varepsilon \in (0, l/2^{m+1})$  and consider for  $i = 1, \dots, M$ , the closed cubes

$$C_i^\varepsilon = \{x \in \mathbb{R}^d : d_2(x, (C_i)^c) \geq \varepsilon\} \subset C_i.$$

Here  $d_2$  denotes the standard Euclidean metric on  $\mathbb{R}^d$ . Now observe that there exists a finite set  $\mathcal{K} = \mathcal{K}(\varepsilon) \subset \mathbb{R}^d$  such that for  $x \in C_i^\varepsilon$ ,  $i = 1, \dots, M$ ,

$$d(x, \mathcal{K} \cap C_i) \leq d(x, (C_i)^c). \quad (13)$$

We extend the codebooks  $\mathcal{C}^{(1)}(N)$  to  $\mathcal{C}^{(2)}(N) = \mathcal{C}^{(1)}(N) \cup \mathcal{K}$  and let  $\mathcal{C}_i(N) = \mathcal{C}^{(2)}(N) \cap C_i^\varepsilon$  for  $N \in \mathbb{I}$  and  $i = 1, \dots, M$ . Note that property (13) guarantees that any point  $x$  in an

arbitrary cube  $C_i^\varepsilon$  has as best  $\mathcal{C}^{(2)}(N)$ -approximant an element in  $\mathcal{C}_i(N)$ . Moreover, none of the codebooks  $\mathcal{C}_i(N)$  is empty, i.e. the number  $N_i = N_i(N)$  defined as  $N_i = |\mathcal{C}_i(N)|$ , is greater or equal to 1. Consequently, letting  $X_i$  denote  $\mathcal{U}(C_i^\varepsilon)$ -distributed r.v.'s, one obtains

$$\begin{aligned} \int \varphi(N^{1/d} d(x, \mathcal{C}^{(2)}(N))) d\mu^m(x) &\geq \int_{\bigcup_{i=1}^M C_i^\varepsilon} \varphi(N^{1/d} d(x, \mathcal{C}^{(2)}(N))) d\mu^m(x) \\ &= \sum_{i=1}^M \mu^m(C_i^\varepsilon) \mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}_i(N))). \end{aligned} \quad (14)$$

Let  $U$  be a  $\mathcal{U}([0, 1]^d)$ -distributed r.v., and fix an arbitrary  $i \in \{1, \dots, M\}$ . Note that the cube  $C_i^\varepsilon$  has side length  $2^{-m}l - 2\varepsilon$ , so that a shifting and rescaling yields

$$\begin{aligned} \mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}_i(N))) &\geq \mathbb{E} \varphi(N^{1/d} (2^{-m}l - 2\varepsilon) \|U - \hat{U}^{(N_i)}\|) \\ &= \mathbb{E} \varphi((N\lambda^d(C_i^\varepsilon))^{1/d} \|U - \hat{U}^{(N_i)}\|), \end{aligned}$$

where  $\hat{U}^{(N_i)}$  denotes an optimal approximation satisfying the range constraint  $|\text{range}(\hat{U}^{(N_i)})| \leq N_i$ . With  $f_N(\eta)$  as in (5), we arrive at

$$\mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}_i(N))) \geq f_{N_i}(N_i / (N\lambda^d(C_i^\varepsilon))).$$

We need to control the quantity  $N_i/N$  for  $N$  large. Recall that  $\mathcal{C}^{(2)}(N)$  is the union of the sets  $\mathcal{C}(N)$ ,  $\mathcal{K}$  and  $(\kappa N^{-1/d} \mathbb{Z}^d) \cap C$ , and the vague convergence of  $\nu^N$  to  $\nu$  implies that

$$\limsup_{N \rightarrow \infty} \frac{|\mathcal{C}(N) \cap C_i|}{N} \leq \nu(\bar{C}_i).$$

Moreover, the set  $(\kappa N^{-1/d} \mathbb{Z}^d) \cap C_i$  contains at most  $(\frac{l 2^{-m}}{\kappa N^{-1/d}} + 1)^d$  elements, so that

$$\limsup_{N \rightarrow \infty} \frac{N_i}{N} \leq \nu(\bar{C}_i) + \frac{\lambda^d(C_i)}{\kappa^d}.$$

Consequently, Theorem 2.1 implies that

$$\mathbb{E} \varphi(N^{1/d} d(X_i, \mathcal{C}_i(N))) \gtrsim g\left(\left(\nu(\bar{C}_i) + \frac{\lambda^d(C_i)}{\kappa^d}\right) / \lambda^d(C_i^\varepsilon)\right).$$

Combining this estimate with (12) and (14) yields

$$\begin{aligned} \int_C \varphi(N^{1/d} d(x, \mathcal{C})) d\mu_c(x) &\geq \int_C \varphi(N^{1/d} d(x, \mathcal{C}^{(1)})) d\mu^m(x) - \varphi(c\kappa) \|h - h_m\|_{L^1(C)} \\ &\gtrsim \sum_{i=1}^M \mu^m(C_i^\varepsilon) g\left(\left(\nu(\bar{C}_i) + \frac{\lambda^d(C_i)}{\kappa^d}\right) / \lambda^d(C_i^\varepsilon)\right) \\ &\quad - \varphi(c\kappa) \|h - h_m\|_{L^1(C)} \end{aligned}$$

as  $N \rightarrow \infty$ . Since  $\varepsilon > 0$  can be chosen arbitrarily small, it follows that

$$\begin{aligned} \int_C \varphi(N^{1/d} d(x, \mathcal{C}^{(2)})) d\mu_c(x) &\gtrsim \sum_{i=1}^M \mu_c(C_i) g\left(\frac{\nu(C_i)}{\lambda^d(C_i)} + \frac{1}{\kappa^d}\right) - \varphi(c\kappa) \|h - h_m\|_{L^1(C)} \\ &= \int_C g\left(\xi_m(x) + \frac{1}{\kappa^d}\right) d\mu_c(x) - \varphi(c\kappa) \|h - h_m\|_{L^1(C)} \end{aligned}$$

As  $m \rightarrow \infty$ , the densities  $\xi_m$  converge pointwise to  $\xi = \frac{d\nu_c}{d\lambda^d}$  for  $\lambda^d$ -a.a.  $x$ , and  $h_m$  converges to  $h$  in  $L^1(C)$ . Consequently, Fatou's Lemma implies that

$$\liminf_{N \rightarrow \infty} \int_C \varphi(N^{1/d} d(x, \mathcal{C}^{(2)})) d\mu_c \geq \int_C g\left(\xi(x) + \frac{1}{\kappa^d}\right) d\mu_c(x).$$

Finally, observing that  $\kappa > 0$  was arbitrary and applying monotone convergence yields the general result.  $\square$

The above proposition enables us to give a partial proof of Theorem 1.3. For the remainder of this section, let  $I$  be given by (3), assume that  $I \in (0, \infty)$ , and denote by  $\mathcal{M}$  the set of measures associated with the minimizers of the point allocation problem as defined in (4). So far we have not proved that  $\mathcal{M}$  is non-empty.

**Proposition 4.2.** *Suppose that the codebooks  $\mathcal{C}(N)$ ,  $N \in \mathbb{N}$ , are of size  $N$  and satisfy*

$$\limsup_{N \rightarrow \infty} N^{1/d} \|d(X, \mathcal{C}(N))\|_\varphi \leq I^{1/d}, \quad (15)$$

*and consider the associated empirical measures*

$$\nu^N = \frac{1}{N} \sum_{\hat{x} \in \mathcal{C}(N)} \delta_{\hat{x}}, \quad N \in \mathbb{N}.$$

*Then  $(\nu^N)_{N \in \mathbb{N}}$  is a tight sequence of probability measures and any accumulation point of  $(\nu^N)$  lies in  $\mathcal{M}$  (in the weak topology).*

**Proof.** Fix an arbitrary vaguely convergent subsequence  $(\nu^N)_{N \in \mathbb{I}}$  of  $(\nu^N)_{N \in \mathbb{N}}$  and denote by  $\nu$  its limiting measure. Let  $\varepsilon > 0$ . As long as the Orlicz norm  $\|d(X, \mathcal{C}(N))\|_\varphi$  is finite, one has in general

$$\mathbb{E} \varphi\left(\frac{d(X, \mathcal{C}(N))}{\|d(X, \mathcal{C}(N))\|_\varphi}\right) \leq 1.$$

Note that (15) implies that for all sufficiently large  $N \in \mathbb{N}$

$$\|d(X, \mathcal{C}(N))\|_\varphi \leq ((1 + \varepsilon)I/N)^{1/d}$$

so that

$$\limsup_{N \rightarrow \infty} \mathbb{E} \varphi((N/(1 + \varepsilon)I)^{1/d} d(X, \mathcal{C}(N))) \leq 1.$$

We consider the codebooks  $\tilde{\mathcal{C}}(\tilde{N}) = \mathcal{C}((1 + \varepsilon)\tilde{N}I)$  for

$$\tilde{N} \in \tilde{\mathbb{I}} := \{N/((1 + \varepsilon)I) : N \in \mathbb{I}\}.$$

Then

$$\limsup_{\tilde{N} \rightarrow \infty} \mathbb{E} \varphi(\tilde{N}^{1/d} d(X, \tilde{\mathcal{C}}(\tilde{N}))) \leq 1. \quad (16)$$

On the other hand, the empirical measures

$$\tilde{\nu}^{\tilde{N}} := \frac{1}{\tilde{N}} \sum_{\tilde{x} \in \tilde{\mathcal{C}}(\tilde{N})} \delta_{\tilde{x}} = (1 + \varepsilon) I \nu^{(1+\varepsilon)\tilde{N}I}$$

converge vaguely to  $(1 + \varepsilon)I\nu$  so that by Theorem 4.1,

$$\liminf_{\tilde{N} \rightarrow \infty} \mathbb{E} \varphi(\tilde{N}^{1/d} d(X, \tilde{\mathcal{C}}(\tilde{N}))) \geq \int g((1 + \varepsilon)I \xi(x)) d\mu_c(x),$$

where  $\xi = \frac{d\nu_c}{d\lambda^d}$ . Combining this with (16), and noticing that  $\varepsilon > 0$  is arbitrary, one obtains

$$\int g(I \xi(x)) d\mu_c(x) \leq 1.$$

Consequently, the point allocation  $\tilde{\xi}(x) = I \xi(x)$  solves the allocation problem:

$$\int g(\tilde{\xi}(x)) d\mu_c(x) \leq 1 \text{ and } \int \tilde{\xi} d\lambda^d \leq I. \quad (17)$$

Due to the definition of  $I$ , the right inequality is actually an equality.

Assume now that  $\int g(\tilde{\xi}(x)) d\mu_c(x) < 1$ , and fix  $\delta > 0$  (small) so that the set  $A := \{x \in \mathbb{R}^d : \tilde{\xi}(x) \geq \delta\}$  has positive Lebesgue measure. Since  $g$  restricted to  $[\delta/2, \infty)$  is Lipschitz continuous, we can lower the density  $\tilde{\xi}$  on  $A$  in such a way that the point allocation constraint remains valid, thus contradicting the optimality of  $I$ . Consequently, the inequalities in (17) are even equalities, and we immediately obtain that  $\nu_c \in \mathcal{M}$ . Since  $\nu_c$  has mass 1, we also have that  $\nu = \nu_c \in \mathcal{M}$ . Moreover,  $(\nu^N)_{N \in \mathbb{N}}$  converges to  $\nu$  in the weak topology. We finish the proof by noticing that the sequence  $(\nu^N)$  is tight, since it has no vaguely convergent subsequence losing some of its mass.  $\square$

## 5 The point allocation problem

We decompose the original measure  $\mu$  into its absolutely continuous part  $\mu_c = h d\lambda^d$  and singular component  $\mu_s$ . The singular component will have no influence on the asymptotics of the quantization error.

In this section we use standard methods for convex optimization problems to treat the point allocation problem, i.e. the minimization of

$$\int_{\mathbb{R}^d} \xi(x) dx \quad (18)$$

over all positive integrable functions  $\xi : \mathbb{R}^d \rightarrow [0, \infty)$  satisfying

$$\int_{\mathbb{R}^d} g(\xi(x)) d\mu_c(x) \leq 1. \quad (19)$$

A minimizer  $\xi$  will be called *optimal point density*.

We shall use the convex conjugate of  $g$ , i.e.

$$g^*(a) = \sup_{\eta \geq 0} [a\eta - g(\eta)], \quad a \leq 0,$$

and the concave function  $\bar{g} : [0, \infty) \rightarrow [0, \infty)$ ,  $a \mapsto -g^*(-a)$ . Alternatively, one can define  $\bar{g}$  as  $\bar{g}(a) = \inf_{\eta \geq 0} [a\eta + g(\eta)]$ .

The function  $\bar{g}$  is continuous and satisfies  $\bar{g}(0) = \inf_{\eta \geq 0} g(\eta) = 0$ . The right continuity in 0 is a consequence of the lower semicontinuity of  $g^*$ . Moreover, since  $g$  is lower semicontinuous, one has

$$g(\eta) = \sup_{a \leq 0} [a\eta - g^*(a)] = \sup_{a \geq 0} [\bar{g}(a) - a\eta], \quad \eta \geq 0. \quad (20)$$

**Theorem 5.1.** 1. *The minimal value  $I$  satisfies the dual formula*

$$I = \sup_{\kappa > 0} \frac{1}{\kappa} \left( \int \bar{g}\left(\frac{\kappa}{h(x)}\right) d\mu_c(x) - 1 \right). \quad (21)$$

2. *The optimization problem has an integrable solution iff the integral*

$$\int \bar{g}\left(\frac{\kappa}{h(x)}\right) d\mu_c(x) \quad (22)$$

*is finite for some  $\kappa > 0$ . In such a case there exists an optimal point density  $\xi$ .*

3. *Suppose that (22) is finite and that*

$$\mu_c(\mathbb{R}^d) \sup_{t \geq 0} \varphi(t) > 1.$$

*Then  $I > 0$  and there exists an optimal point density. Moreover, all optimal point densities  $\xi$  satisfy*

$$\int g(\xi(x)) d\mu_c(x) = 1 \quad \text{and} \quad \bar{g}'_+\left(\frac{\kappa}{h(x)}\right) \leq \xi(x) \leq \bar{g}'_-\left(\frac{\kappa}{h(x)}\right) \text{ for a.e. } x \in \mathbb{R}^d, \quad (23)$$

*where  $\kappa$  is a maximizer of the right hand side of (21). (Here we make use of the convention that  $\bar{g}'_+(\infty) = \bar{g}'_-(\infty) = 0$ ). In particular, the supremum in the dual formula is attained.*

4. *If  $\mu_c(\mathbb{R}^d) \sup_{t \geq 0} \varphi(t) \leq 1$ , then  $I = 0$  and  $\xi = 0$  is an optimal point density.*

**Remark 5.2.** Assume that  $I \in (0, \infty)$  and that  $g$  is strictly convex. A standard result from convex analysis (see Rockafellar (1970), Theorem 26.3) states that the strict convexity is equivalent to differentiability of  $\bar{g}$ . Hence, one obtains a one parameter family of candidates as optimal point densities. Moreover,

$$\bar{g}'(a) = \inf\{b > 0 : -g'_+(b) \leq a\}. \quad (24)$$

Additionally, the strict convexity implies almost everywhere uniqueness, since for two optimal solutions  $\xi_1, \xi_2$  that were not almost everywhere identical the combination  $\bar{\xi} = \frac{1}{2}(\xi_1 + \xi_2)$  would satisfy  $\int g(\bar{\xi}(x)) dx < 1$  and by continuity of  $g$  it is straight forward to construct an admissible density with smaller  $L^1$ -norm.

If  $g$  is additionally differentiable, the one parameter family of candidates is given via

$$\xi^\kappa(x) = (-g')^{-1}\left(\frac{\kappa}{h(x)}\right)$$

with the convention that  $(-g')^{-1}(\infty) = 0$ .

**Proof of Theorem 5.1.** By the concavity of  $\bar{g}$ , the integral (22) is either finite or infinite for all  $\kappa > 0$ . We start with proving the “ $\geq$ ” inequality in the dual formula. Note that by definition of  $\bar{g}$ ,  $a b \geq \bar{g}(a) - g(b)$  for  $a, b \geq 0$ . Therefore, for  $\kappa > 0$  and  $\xi$  satisfying (19), it is true that

$$\begin{aligned} \int \xi(x) dx &\geq \int_{\{h>0\}} \xi(x) dx = \frac{1}{\kappa} \int \frac{\kappa}{h(x)} \xi(x) d\mu_c(x) \\ &\geq \frac{1}{\kappa} \left( \int \bar{g}\left(\frac{\kappa}{h(x)}\right) d\mu_c(x) - \int g(\xi(x)) d\mu_c(x) \right) \geq \frac{1}{\kappa} \left( \int \bar{g}\left(\frac{\kappa}{h(x)}\right) d\mu_c(x) - 1 \right). \end{aligned} \quad (25)$$

In order to have equalities in the above estimates, we need to find a density  $\xi$  and  $\kappa > 0$  such that

$$\xi(x) = 0 \text{ for } \lambda^d \text{ a.a. } x \in \{h = 0\}, \quad (26)$$

$$\bar{g}\left(\frac{\kappa}{h(x)}\right) - g(\xi(x)) = \frac{\kappa}{h(x)} \xi(x), \quad \text{for } \mu_c \text{ a.a. } x \quad (27)$$

and

$$\int g(\xi(x)) d\mu_c(x) = 1. \quad (28)$$

In the case  $\mu_c(\mathbb{R}^d) \sup_{t \geq 0} \varphi(t) \leq 1$ , it is easily seen that  $\xi = 0$  is an optimal point density so that  $I = 0$  which proves assertion 4. Moreover, the term on the right hand side of (21) tends to 0 when letting  $\kappa \rightarrow \infty$  so that the dual formula is valid in that case. Moreover, if (5) is infinite for one  $\kappa > 0$ , then there is no integrable nonnegative function  $\xi$  satisfying (19) and the dual formula is valid as well.

From now on, we assume that  $\mu_c(\mathbb{R}^d) \sup_{t \geq 0} \varphi(t) > 1$  and that (5) is finite for any  $\kappa > 0$ . Next, we derive a density  $\xi$  satisfying the three abovementioned conditions. Then estimate (25) implies optimality for this choice of  $\xi$  which proves assertion 1.

First we examine the second condition. Consider  $a_0, b_0 \geq 0$  with

$$\bar{g}'_+(a_0) \leq b_0 \leq \bar{g}'_-(a_0). \quad (29)$$

Due to the concavity of  $\bar{g}$  it holds that  $\bar{g}(a) \leq \bar{g}(a_0) + b_0(a - a_0)$  for all  $a \geq 0$ , and we obtain with (20),

$$g(b_0) = \sup_{a \geq 0} [\bar{g}(a) - b_0 a] \leq \sup_{a \geq 0} [\bar{g}(a_0) + b_0(a - a_0) - b_0 a] = \bar{g}(a_0) - b_0 a_0.$$

Consequently, condition (29) implies that

$$g(b_0) = \bar{g}(a_0) - b_0 a_0. \quad (30)$$

Conversely, it is easy to see that any pair  $(a_0, b_0)$  of nonnegative reals satisfying (30) also satisfy (29).

For  $\kappa > 0$ , we consider the point densities

$$\xi_+^\kappa(x) = \bar{g}'_+\left(\frac{\kappa}{h(x)}\right), \quad x \in \mathbb{R}^d,$$

and

$$\xi_-^\kappa(x) = \bar{g}'_-\left(\frac{\kappa}{h(x)}\right), \quad x \in \mathbb{R}^d$$

with the convention  $\bar{g}'_+(\infty) = \bar{g}'_-(\infty) = 0$  so that, in particular,  $\xi_-^\kappa(x) = \xi_+^\kappa(x) = 0$  for  $x \in \{h = 0\}$ . Furthermore, for any  $x \in \{h > 0\}$ , condition (29) is satisfied for  $a_0 = \kappa/h(x)$  and for all  $b_0 \in [\xi_+^\kappa(x), \xi_-^\kappa(x)]$ . Therefore, every convex combination  $\bar{\xi} = \alpha \xi_+^\kappa + (1 - \alpha) \xi_-^\kappa$  satisfies

$$\bar{g}\left(\frac{\kappa}{h(x)}\right) - g(\bar{\xi}(x)) = \frac{\kappa}{h(x)} \bar{\xi}(x) \quad \text{on } \{h > 0\}. \quad (31)$$

In analogy to (25) we obtain that

$$\int \bar{\xi}(x) dx = \frac{1}{\kappa} \left( \int \bar{g}\left(\frac{\kappa}{h(x)}\right) d\mu_c(x) - \int g(\bar{\xi}(x)) d\mu_c(x) \right).$$

In particular, all three integrals are finite. It remains to find an appropriate  $\kappa > 0$  and a convex combination  $\bar{\xi}$  as above with

$$\int g(\bar{\xi}(x)) d\mu_c(x) = 1.$$

We need to compute the asymptotic behavior of  $g(\bar{g}'_-(a))$  as  $a \rightarrow 0$  and  $a \rightarrow \infty$ . Due to equation (30), one has for all  $a > 0$ ,

$$g(\bar{g}'_-(a)) = \bar{g}(a) - \bar{g}'_-(a) a \leq \bar{g}(a),$$

and we obtain that  $\lim_{a \downarrow 0} g(\bar{g}'_-(a)) = 0$ . On the other hand, for any  $b > 0$ ,

$$g(b) = \sup_{a \geq 0} [\bar{g}(a) - a b] \geq \limsup_{a \rightarrow \infty} a \left( \frac{\bar{g}(a)}{a} - b \right),$$

and, since  $g(b)$  is finite, it follows that  $\lim_{a \rightarrow \infty} \bar{g}(a)/a = 0$ . Thus the concavity of  $\bar{g}$  implies that  $\lim_{a \rightarrow \infty} \bar{g}'_-(a) = 0$  and we arrive at

$$\lim_{a \rightarrow \infty} g(\bar{g}'_-(a)) = g(0).$$

The above asymptotics imply with monotone convergence, that

$$\lim_{\kappa \rightarrow 0} \int g(\xi_-^\kappa(x)) d\mu_c(x) = 0,$$

and

$$\lim_{\kappa \rightarrow \infty} \int g(\xi_-^\kappa(x)) d\mu_c(x) = g(0) \mu_c(\mathbb{R}^d) > 1.$$

Now let

$$\kappa_0 = \sup \left\{ \kappa > 0 : \int g(\xi_-^\kappa(x)) d\mu_c(x) \leq 1 \right\}.$$

As we have seen above the set is non-empty and bounded so that  $\kappa_0 \in (0, \infty)$ . Moreover, since for any  $x \in \mathbb{R}^d$ ,  $\kappa \mapsto \xi_-^\kappa(x)$  is decreasing and left continuous, it follows by monotone convergence that

$$\int g(\xi_-^{\kappa_0}(x)) d\mu_c(x) = \lim_{\kappa \uparrow \kappa_0} \int g(\xi_-^\kappa(x)) d\mu_c(x) \leq 1. \quad (32)$$

Similarly the inequality  $\xi_+^\kappa(x) \leq \xi_-^\kappa(x)$  implies that

$$\begin{aligned} \int g(\xi_+^{\kappa_0}(x)) d\mu_c(x) &= \lim_{\kappa \downarrow \kappa_0} \int g(\xi_+^\kappa(x)) d\mu_c(x) \\ &\geq \lim_{\kappa \downarrow \kappa_0} \int g(\xi_-^\kappa(x)) d\mu_c(x) \geq 1. \end{aligned} \quad (33)$$

Hence, inequalities (32) and (33) imply the existence of a convex combination  $\bar{\xi} = \alpha \xi_+^\kappa + (1 - \alpha) \xi_-^\kappa$  with

$$\int g(\bar{\xi}(x)) d\mu_c(x) = 1.$$

This function  $\bar{\xi}$  solves

$$\int \bar{\xi}(x) dx = \frac{1}{\kappa_0} \left( \int \bar{g}\left(\frac{\kappa_0}{h(x)}\right) d\mu_c(x) - 1 \right) = \sup_{\kappa > 0} \frac{1}{\kappa} \left( \int \bar{g}\left(\frac{\kappa}{h(x)}\right) d\mu_c(x) - 1 \right),$$

and we proved assertion 1. Moreover, we observe that for any  $\xi$  which is not of the form (23) there is a strict inequality in at least one of the estimates in (25).  $\square$

## 6 The singular case

In this section, we consider an original  $X$  with law  $\mu \perp \lambda^d$  or, equivalently,  $\mu_c = 0$ . Moreover, we again assume that  $\mu$  is compactly supported.

**Proposition 6.1.** *There exist codebooks  $\mathcal{C}(N)$ ,  $N \geq 1$ , with  $\lim_{N \rightarrow \infty} \frac{1}{N} |\mathcal{C}(N)| = 0$  such that*

$$\lim_{N \rightarrow \infty} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) = 0.$$

**Proof.** We fix  $\varepsilon > 0$ . As the measure  $\mu$  is singular with respect to the Lebesgue measure  $\lambda^d$ , there exists an open set  $A \subset \mathbb{R}^d$  with  $\mu(A) = 1$  and  $\lambda^d(A) \leq \varepsilon$ . Due to Lemma 1.4.2 in Cohn (1980), one can represent the open set  $A$  as a countable disjoint union of half-open cubes  $(C_i)_{i \in \mathbb{N}}$  in  $\bigcup_{m=1}^{\infty} \mathcal{B}_m$ , where

$$\mathcal{B}_m = \{[i_1 2^{-m}, (i_1 + 1) 2^{-m}) \times \cdots \times [i_d 2^{-m}, (i_d + 1) 2^{-m}) : i_1, \dots, i_d \in \mathbb{Z}\} \subset \mathbb{R}^d$$

for  $m \in \mathbb{N}$ . Due to monotone convergence, we obtain that there exists  $M \in \mathbb{N}$  with

$$\mu\left(\bigcup_{i=1}^M C_i\right) \geq \mu(A) - \varepsilon = 1 - \varepsilon. \quad (34)$$

Set  $C = \bigcup_{i=1}^M C_i$ .

Let us introduce the codebooks; fixing  $l > 0$  such that  $\text{supp}(\mu) \subset [-l, l]^d$ , the construction depends upon two parameters  $\kappa_1, \kappa_2 > 0$ :

$$\mathcal{C}(N) = ((\kappa_1 N^{-1/d} \mathbb{Z}^d) \cap C) \cup ((\kappa_2 N^{-1/d} \mathbb{Z}^d) \cap [-l, l]^d), \quad N \geq 1.$$

We need to control the size of  $\mathcal{C}(N)$ . For  $i \in \{1, \dots, M\}$ , let  $m_i \in \mathbb{N}$  denote the unique number with  $C_i \in \mathcal{B}_{m_i}$ , and observe that

$$|(\kappa_1 N^{-1/d} \mathbb{Z}^d) \cap C_i| \leq \left(\frac{2^{-m_i}}{\kappa_1 N^{-1/d}} + 1\right)^d \sim \lambda^d(C_i) \kappa_1^{-d} N$$

as  $N \rightarrow \infty$ . Analogously,

$$|(\kappa_2 N^{-1/d} \mathbb{Z}^d) \cap [-l, l]^d| \leq \left(\frac{2l}{\kappa_2 N^{-1/d}} + 1\right)^d \sim (2l)^d \kappa_2^{-d} N.$$

Consequently,

$$\begin{aligned} |\mathcal{C}(N)| &\leq \sum_{i=1}^M |(\kappa_1 N^{-1/d} \mathbb{Z}^d) \cap C_i| + |(\kappa_2 N^{-1/d} \mathbb{Z}^d) \cap [-l, l]^d| \\ &\lesssim (\lambda^d(C) \kappa_1^{-d} + (2l)^d \kappa_2^{-d}) N \leq (\varepsilon \kappa_1^{-d} + (2l)^d \kappa_2^{-d}) N. \end{aligned}$$

Next, we estimate the approximation error. Suppose that  $N \geq 1$  is sufficiently large so that  $C_i \cap \mathcal{C}(N) \neq \emptyset$  for all  $i = 1, \dots, M$ . Let  $c = \sup_{x \in [0,1]^d} \|x\|$ , and observe that for all

$x \in C$ ,  $d(x, \mathcal{C}(N)) \leq c\kappa_1 N^{-1/d}$ . Moreover, for any  $x \in [-l, l]^d$ ,  $d(x, \mathcal{C}(N)) \leq c\kappa_2 N^{-1/d}$ . Consequently,

$$\begin{aligned} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) &\leq \mu(C) \varphi(c\kappa_1) + (1 - \mu(C)) \varphi(c\kappa_2) \\ &\leq \varphi(c\kappa_1) + \varepsilon \varphi(c\kappa_2). \end{aligned} \quad (35)$$

Now, for  $\delta > 0$  arbitrary, pick  $\kappa_1, \kappa_2 > 0$  satisfying  $\varphi(c\kappa_1) \leq \delta/2$  and  $(2l)^d/\kappa_2^d \leq \delta/2$ , and choose  $\varepsilon > 0$  so that  $\varepsilon\kappa_1^{-d} \leq \delta/2$  and  $\varepsilon \varphi(c\kappa_2) \leq \delta/2$ . Then the corresponding codebooks  $\mathcal{C}(N)$  satisfy

$$|\mathcal{C}(N)| \lesssim \delta N \quad \text{and} \quad \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) \lesssim \delta,$$

and the assertion of the proposition follows by a diagonalization argument.  $\square$

## 7 Extension to the non-compact setting

In order to treat the non-compact quantization problem, we need to control the impact of realizations lying outside large cubes. For  $L^p(\mathbb{P})$ -norm distortions, Pierce (1970) (see also Graf and Luschgy (2000), Lemma 6.6) discovered that the quantization error can be estimated against a higher moment  $\tilde{p} > p$  of  $\|X\|$ . His result can be easily extended to the inequality

$$\delta(N|X, p) \leq C \mathbb{E}[\|X\|^{\tilde{p}}]^{1/\tilde{p}} N^{-1/d},$$

where  $X$  is an arbitrary original in  $\mathbb{R}^d$ ,  $N \in \mathbb{N}$  and  $C$  is a universal constant depending only on  $E$ ,  $p$  and  $\tilde{p}$ . Pierce's proof is based on a random coding argument. In contrast to his approach, we will use  $\varepsilon$ -nets to establish a similar result.

The construction is based on several parameters. Let  $\Psi : [0, \infty) \rightarrow [0, \infty)$  denote an increasing function,  $(r_n)_{n \in \mathbb{N}_0}$  an increasing sequence, and let  $(\alpha_n)_{n \in \mathbb{N}}$  be a positive decreasing and summable sequence.

**Lemma 7.1.** *Let  $J \in \mathbb{N}_0$  and denote by  $X$  a  $(B(0, r_J)^c \cup \{0\})$ -valued r.v. For  $N \geq 0$  there exists a codebook  $\mathcal{C}(N)$  of size  $1 + N \sum_{n=J}^{\infty} \alpha_n$  satisfying*

$$\mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C}(N))) \leq \mathbb{E}[\Psi(\|X\|)] \sum_{n=J}^{\infty} \frac{1}{\Psi(r_n)} \varphi(c_E \alpha_n^{-1/d} r_{n+1}), \quad (36)$$

where  $c_E$  is a finite constant depending on the norm  $\|\cdot\|$  only.

**Proof.** First, observe that the sum in estimate (36) diverges whenever  $\liminf_{n \rightarrow \infty} r_n < \infty$ . Thus we can assume without loss of generality that  $\lim_{n \rightarrow \infty} r_n = \infty$ . Fix  $J \in \mathbb{N}_0$  and  $N \in \mathbb{N}$ . For  $n \in \mathbb{N}_0$ , let  $V_n = B(0, r_n)$  and  $N_n = \alpha_n N$ . Moreover, we denote by  $I \in \mathbb{N}_0$  the smallest index  $n$  with  $N_n < 1$ , and let for  $n \in \mathbb{N}_0$  with  $J \leq n < I$ ,  $\mathcal{C}_n$  denote an optimal  $\varepsilon$ -net for  $B(0, r_{n+1})$  consisting of  $N_n$  elements. As is well known there exists a constant  $c_E$  only depending on the norm  $\|\cdot\|$  such that  $d(x, \mathcal{C}_n) \leq c_E r_{n+1} N_n^{-1/d}$  for all  $x \in B(0, r_{n+1})$ .

Thus the codebook  $\mathcal{C} = \{0\} \cup \bigcup_{n=J}^{I-1} \mathcal{C}_n$  contains at most  $1 + N \sum_{n=J}^{\infty} \alpha_n$  elements, and we have

$$\begin{aligned} \mathbb{E} \varphi(N^{1/d} d(X, \mathcal{C})) &= \sum_{n=J}^{\infty} \mathbb{E} \left[ 1_{V_{n+1} \setminus V_n}(X) \varphi(N^{1/d} d(X, \mathcal{C})) \right] \\ &\leq \sum_{n=J}^{\infty} \mathbb{P}(X \notin V_n) \varphi(c_E r_{n+1} N^{1/d} / (1 \vee N_n)^{1/d}) \\ &\leq \sum_{n=J}^{\infty} \mathbb{P}(X \notin V_n) \varphi(c_E r_{n+1} \alpha_n^{-1/d}) \\ &\leq \mathbb{E}[\Psi(\|X\|)] \sum_{n=J}^{\infty} \frac{1}{\Psi(r_n)} \varphi(c_E r_{n+1} \alpha_n^{-1/d}). \end{aligned}$$

□

**Definition 7.2.** We say that an increasing function  $\Psi : [0, \infty) \rightarrow [0, \infty)$  satisfies the growth condition (G) for  $\varphi$  iff there exists a decreasing summable sequence  $(\alpha_n)$  and an increasing sequence  $(r_n)$  with

$$\sum_{n=0}^{\infty} \frac{1}{\Psi(r_n)} \varphi(\alpha_n^{-1/d} r_{n+1}) < \infty.$$

Suppose that  $\Psi$  satisfies condition (G) for  $\varphi$ . We shall see that the condition that  $\mathbb{E}\Psi(\|X\|) < \infty$  is sufficient to conclude that the quantization problem for  $X$  under the Orlicz norm induced by  $\varphi$  is of order  $N^{-1/d}$ . Moreover, quantization of non-compact measures then can be approximated by the compact setting.

**Example 7.3.** • Suppose that  $\varphi(t) \leq c(1 + t^p)$  for all  $t \geq 0$ , where  $c, p \in \mathbb{R}_+$  are appropriate constants. Then, for any  $\beta > \frac{p+d}{d}$ ,

$$\Psi(t) = t^p (\log_+ t)^\beta$$

satisfies (G) for  $\varphi$ . For instance, one may choose  $(r_n)_{n \in \mathbb{N}_0} = (2^n)_{n \in \mathbb{N}_0}$  and  $(\alpha_n)_{n \in \mathbb{N}_0} = ((n+2)^{-\gamma})_{n \in \mathbb{N}_0}$  for a  $\gamma \in (1, \frac{d}{p}(\beta-1))$ .

- Suppose that  $\varphi$  satisfies  $\varphi(t) \leq c \exp\{t^\kappa\}$  for all  $t \geq 0$ , where  $c, \kappa \in \mathbb{R}_+$  are appropriate constants. Then, for any  $\tilde{\kappa} > \kappa$ , the function

$$\Psi(t) = \exp\{t^{\tilde{\kappa}}\}$$

satisfies (G) for  $\varphi$ , as can be verified easily for  $(\alpha_n)_{n \in \mathbb{N}_0} = ((n+2)^{-2})_{n \in \mathbb{N}_0}$  and  $(r_n)_{n \in \mathbb{N}_0} = ((n+1)^s)_{n \in \mathbb{N}_0}$  for  $s > 0$  with  $s\tilde{\kappa} > (\frac{2}{d} + s)\kappa$ .

**Remark 7.4.** The proof of the upper bound in Theorem 1.1 relies on the assumption that  $\mathbb{E}\Psi(\|X\|) < \infty$  for some  $\Psi$  satisfying the growth condition (G). As we shall see below, this assumption can be replaced by the equivalent condition that  $X \in L^\Psi(\mathbb{P})$  for

some  $\Psi$  satisfying (G). First, assume that  $\mathbb{E}\Psi(\|X\|) < \infty$  for some  $\Psi$  satisfying (G). Then  $\tilde{\Psi} = 1_{[1, \infty)} \Psi$  satisfies (G), and since by monotone convergence

$$\lim_{\kappa \rightarrow \infty} \mathbb{E}\Psi(\|X\|/\kappa) = 0,$$

it follows that  $X \in L^{\tilde{\Psi}}(\mathbb{P})$ . On the other hand, assuming that  $X \in L^{\Psi}(\mathbb{P})$  for some  $\Psi$  satisfying (G) implies the existence of a  $\kappa > 0$  for which

$$\mathbb{E}\Psi(\|X\|/\kappa) < \infty.$$

Now, let  $\tilde{\Psi}(t) = \Psi(t/\kappa)$ , and denote by  $(\alpha_n)$  and  $(r_n)$  sequences as in Definition 7.2. Then  $\mathbb{E}\tilde{\Psi}(\|X\|) < \infty$ , and

$$\sum_{n=0}^{\infty} \frac{1}{\tilde{\Psi}(\tilde{r}_n)} \varphi(\tilde{\alpha}_n^{-1/d} \tilde{r}_{n+1}) = \sum_{n=0}^{\infty} \frac{1}{\Psi(r_n)} \varphi(\alpha_n^{-1/d} r_{n+1}) < \infty$$

for  $\tilde{\alpha}_n = \kappa^d \alpha_n$  and  $\tilde{r}_n = \kappa r_n$ ,  $n \in \mathbb{N}_0$ .

We now combine the quantization results for continuous, singular, and unbounded measures to finish the proof of the upper bound in Theorem 1.1.

**Proof of Theorem 1.1.** Let  $\Psi$  be a function satisfying (G). It is easy to see that there exist also a summable and decreasing sequence  $(\alpha_n)$  and an increasing sequence  $(r_n)$  such that

$$\sum_{n=0}^{\infty} \frac{1}{\Psi(r_n)} \varphi(c_E \alpha_n^{-1/d} r_{n+1}) < \infty,$$

where  $c_E$  is as in Lemma 7.1. We denote by  $\xi$  an optimal point density, so that  $\xi$  satisfies

$$\int_{\mathbb{R}^d} g(\xi(x)) d\mu_c(x) = 1 \quad \text{and} \quad \int_{\mathbb{R}^d} \xi(x) dx = I$$

or  $\xi = 0$  (in the case  $I = 0$ ). We fix  $\varepsilon > 0$  and let  $\tilde{\xi}(x) = \xi(x) + \varepsilon h(x)$ ,  $x \in \mathbb{R}^d$ . This point density satisfies

$$\int_{\mathbb{R}^d} g(\tilde{\xi}(x)) d\mu_c(x) < 1$$

since  $g$  is strictly decreasing on  $\{\eta > 0 : g(\eta) > 0\}$ . Now fix  $J \in \mathbb{N}_0$  such that

$$\int_{\mathbb{R}^d} g(\tilde{\xi}(x)) d\mu_c(x) + \mathbb{E}[\Psi(\|X\|)] \sum_{n=J}^{\infty} \frac{1}{\Psi(r_n)} \varphi(c_E \alpha_n^{-1/d} r_{n+1}) < 1 \quad (37)$$

and  $\sum_{n=J}^{\infty} \alpha_n < \varepsilon$ . Next, decompose the measure  $\mu$  into the sum  $\mu = \tilde{\mu}_c + \tilde{\mu}_s + \mu_u$ , where  $\tilde{\mu}_c$  and  $\tilde{\mu}_s$  are the absolutely continuous and singular part of  $\mu$  restricted to  $B(0, r_J)$ , respectively, and  $\mu_u$  contains the rest of the mass of  $\mu$ .

It remains to combine the former results. Due to Proposition 3.1 there exist codebooks  $\mathcal{C}_1(N)$ ,  $N \geq 1$ , with  $\lim_{N \rightarrow \infty} \frac{1}{N} |\mathcal{C}_1(N)| = \|\tilde{\xi}\|_{L^1(\mathbb{R}^d)}$  and

$$\limsup_{N \rightarrow \infty} \int \varphi(N^{1/d} d(x, \mathcal{C}_1(N))) d\tilde{\mu}_c \leq \int_{\mathbb{R}^d} g(\tilde{\xi}(x)) d\mu_c(x).$$

Moreover, Proposition 6.1 implies the existence of codebooks  $\mathcal{C}_2(N)$ ,  $N \geq 1$ , with  $\lim_{N \rightarrow \infty} \frac{1}{N} |\mathcal{C}_2(N)| = 0$  and

$$\lim_{N \rightarrow \infty} \int \varphi(N^{1/d} d(x, \mathcal{C}_2(N))) d\tilde{\mu}_s = 0.$$

Finally, Lemma 7.1 (applied to  $\tilde{X} = 1_{B(0, r_J)^c}(X) \cdot X$ ) yields the existence of codebooks  $\mathcal{C}_3(N)$ ,  $N \geq 1$ , for which

$$\limsup_{N \rightarrow \infty} \frac{1}{N} |\mathcal{C}_3(N)| \leq \sum_{n=J}^{\infty} \alpha_n < \varepsilon$$

and

$$\limsup_{N \rightarrow \infty} \int \varphi(N^{1/d} d(x, \mathcal{C}_3(N))) d\mu_u \leq \mathbb{E}[\Psi(\|X\|)] \sum_{n=J}^{\infty} \frac{1}{\Psi(r_n)} \varphi(C \alpha_n^{-1/d} r_{n+1}).$$

Now consider the codebooks  $\mathcal{C}(N) = \mathcal{C}_1(N) \cup \mathcal{C}_2(N) \cup \mathcal{C}_3(N)$ . Due to the above estimates and (37), one has

$$\begin{aligned} \limsup_{N \rightarrow \infty} \int \varphi(N^{1/d} d(x, \mathcal{C}(N))) d\mu \\ \leq \int_{\mathbb{R}^d} g(\tilde{\xi}(x)) d\mu_c(x) + \mathbb{E}[\Psi(\|X\|)] \sum_{n=J}^{\infty} \frac{1}{\Psi(r_n)} \varphi(C \alpha_n^{-1/d} r_{n+1}) < 1, \end{aligned}$$

so that for sufficiently large  $N$  it is true that  $\|d(X, \mathcal{C}(N))\|_{\varphi} \leq N^{-1/d}$ . On the other hand,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} |\mathcal{C}(N)| < (1 + \varepsilon)I + \varepsilon$$

and, for sufficiently large  $N$ , it holds that  $|\mathcal{C}(N)| \leq (I + \varepsilon I + \varepsilon)N$ . Consequently, it follows that for large  $N$

$$\delta((I + \varepsilon I + \varepsilon)N | X, \varphi) \leq N^{-1/d}.$$

Switching from  $N$  to  $M = (I + \varepsilon I + \varepsilon)N$  one obtains

$$\delta(M | X, \varphi) \leq (I + \varepsilon I + \varepsilon)^{1/d} M^{-1/d},$$

for  $M$  large. Since  $\varepsilon > 0$  was arbitrary, it follows that

$$\limsup_{M \rightarrow \infty} M^{1/d} \delta(M | X, \varphi) \leq I^{1/d}$$

and we proved the upper inequality.

In order to prove the lower bound we fix codebooks  $\mathcal{C}(N)$ ,  $N \in \mathbb{N}$ , with at most  $N$  elements and

$$\limsup_{N \rightarrow \infty} N^{1/d} \delta(N | X, \varphi) \leq I^{1/d}.$$

By Proposition 4.2, each accumulation point of the associated empirical measures

$$\nu^N = \frac{1}{N} \sum_{\hat{x} \in \mathcal{C}(N)} \delta_{\hat{x}}, \quad N \in \mathbb{N},$$

lies in  $\mathcal{M}$ . In particular,

$$\lim_{N \rightarrow \infty} \frac{|\mathcal{C}(N)|}{N} = 1. \quad (38)$$

Therefore, for any  $\varepsilon \in (0, 1)$ ,

$$\liminf_{N \rightarrow \infty} N^{1/d} \delta((1 - \varepsilon)N | X, \varphi) > I^{1/d}.$$

Otherwise one could construct a sequence of codebooks  $\mathcal{C}(N)$ ,  $N \in \mathbb{N}$ , as above which does not fulfil (38). Switching from  $N$  to  $M = (1 - \varepsilon)N$  and letting  $\varepsilon \downarrow 0$  we obtain the lower bound.

The remaining properties of the minimizer  $I$  were proved in Theorem 5.1.  $\square$

## References

- Billingsley, P. 1979. *Probability and measure*. Wiley Series in Probability and mathematical Statistics.
- Bucklew, J. A. 1984. “Two results on the asymptotic performance of quantizers.” *IEEE Trans. Inform. Theory* 30(2, part 2):341–348.
- Bucklew, J. A. and G. L. Wise. 1982. “Multidimensional asymptotic quantization theory with  $r$ th power distortion measures.” *IEEE Trans. Inf. Theory* 28:239–247.
- Cohn, D. L. 1980. *Measure theory*. Mass.: Birkhäuser Boston.
- Cover, T. M. and J. A. Thomas. 1991. *Elements of information theory*. Wiley Series in Telecommunications. New York: John Wiley & Sons, Inc.
- Delattre, S., S. Graf, H. Luschgy and G. Pagès. 2004. “Quantization of probability distributions under norm-based distortion measures.” *Statist. Decision* 22:261–282.
- Dereich, S., F. Fehringer, A. Matoussi and M. Scheutzow. 2003. “On the link between small ball probabilities and the quantization problem for Gaussian measures on Banach spaces.” *J. Theoret. Probab.* 16(1):249–265.
- Fejes Tóth, L. 1972. *Lagerungen in der Ebene, auf der Kugel und im Raum*. 2nd ed. Springer-Verlag.
- Gersho, A. 1979. “Asymptotically optimal block quantization.” *IEEE Trans. Inform. Theory* 25(4):373–380.
- Gersho, A. and R. M. Gray. 1992. *Vector quantization and signal compression*. Boston, MA: Kluwer Academic Publishers.
- Graf, S. and H. Luschgy. 2000. *Foundations of quantization for probability distributions*. Lecture Notes in Mathematics 1730, Berlin: Springer.

- Gray, R. M. and D. L. Neuhoff. 1998. “Quantization.” *IEEE Trans. Inf. Theory* 44(6):2325–2383.
- Gruber, P. M. 2004. “Optimum quantization and its applications.” *Adv. Math.* 186(2):456–497.
- Luschgy, H. and G. Pagès. 2004. “Sharp asymptotics of the functional quantization problem for Gaussian processes.” *Ann. Probab.* 32(2):1574–1599.
- Pagès, Gilles, Huyên Pham and Jacques Printems. 2004. Optimal quantization methods and applications to numerical problems in finance. In *Handbook of computational and numerical methods in finance*. Boston, MA: Birkhäuser Boston pp. 253–297.
- Pierce, J. N. 1970. “Asymptotic quantizing error for unbounded random variables.” *IEEE Trans. Inf. Theory* 16:81–83.
- Rockafellar, R. T. 1970. *Convex analysis*. 2nd ed. Princeton, N. J.
- Zador, P. L. 1966. “Topics in the asymptotic quantization of continuous random variables.” Bell Laboratories Technical Memorandum.
- Zador, P. L. 1982. “Asymptotic quantization error of continuous signals and the quantization dimension.” *IEEE Trans. Inf. Theory* 28:139–149.